Microphone Array Signal Processing for Robot Audition

Heinrich Löllmann, Alastair Moore, Patrick Naylor, Boaz Rafaely, Radu Horaud, Alexandre Mazel, Walter Kellermann

MICROPHONE ARRAY SIGNAL PROCESSING FOR ROBOT AUDITION

Heinrich W. Löllmann[1)], *Alastair H. Moore*[2)], *Patrick A. Naylor*[2)], *Boaz Rafaely*[3)], *Radu Horaud*[4)], *Alexandre Mazel*[5)], *and Walter Kellermann*[1)]

[1)]Friedrich-Alexander University Erlangen-Nürnberg, [2)]Imperial College London, [3)]Ben-Gurion University of the Negev, [4)]INRIA Grenoble Rhône-Alpes, [5)]Softbank Robotics Europe

ABSTRACT

Robot audition for humanoid robots interacting naturally with humans in an unconstrained real-world environment is a hitherto unsolved challenge. The recorded microphone signals are usually distorted by background and interfering noise sources (speakers) as well as room reverberation. In addition, the movements of a robot and its actuators cause ego-noise which degrades the recorded signals significantly. The movement of the robot body and its head also complicates the detection and tracking of the desired, possibly moving, sound sources of interest. This paper presents an overview of the concepts in microphone array processing for robot audition and some recent achievements.

***Index Terms*—** Humanoid robots, robot audition, microphone array processing, ego-noise suppression, source tracking

1. INTRODUCTION

Developing a humanoid robot, which can interact with humans in a natural, i.e., *humanoid* way, is a long-lasting vision of scientists, and with the availability of increasingly powerful technologies, it turns into a realistic engineering task. With the acoustic domain as key modality for voice communication, scene analysis and understanding, acoustic signal processing represents one of the main avenues leading to a humanoid robot, but has received significantly less attention in the past decades than processing in the visual domain.

The design of a system for *robot audition*, which should be operated in real-world environments, starts with the observation that the recorded microphone signals are typically impaired by background and interfering noise sources (speakers) as well as room reverberation [1, 2]. Thereby, the distance between robot and speaker is relatively large in comparison to, e.g., hands-free communication systems for mobile phones. In addition, the movements of a robot, its actuators (motors) and CPU cooling fan cause ego-noise (self-noise) which degrades the recorded signals significantly. Not at least, the movements of the robot body and its head also complicate the detection and tracking of the desired, possibly moving, speaker(s), cf., [3]. Finally, the implementation of algorithms on a robot is often linked to mundane hardware-related problems: The microphone, video and motor data streams are not necessarily synchronized. Besides, the interaction with a robot requires real-time processing where the limited CPU power of an autonomous robot precludes algorithms with a high computational load. A high algorithmic signal delay cannot be allowed either, as a humanoid robot should react, similar to humans, instantaneously to acoustic events in its environment.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 609465. It has been conducted as part of the project EARS (Embodied Audition for RobotS).

There are different concepts and platforms for robot audition, e.g., [1, 4]. The block diagram of Fig. 1 shows a possible realization for a robot audition system, where the components for microphone array processing are marked by gray color. Such a system has been considered within the EU-funded project *Embodied Audition for RobotS* (EARS) whose goal was to develop new algorithms for a natural interaction between humans and robots by means of speech communication.[1]

The "relevant" robot sensing is performed by its microphones and cameras, whose data are used for audio and visual localization and tracking. The microphones are usually embedded in the robot head, but might also be mounted at its body or limbs. The estimates of the direction of arrival (DOA), obtained by (joint) audio and visual tracking, are fed into the attention system of the robot (cf., [5]) where the desired speaker might be identified with support of the dialogue system. The attention system can also be used to control the robot movements based on the speech dialogue (e.g., the robot turns its heads towards the target speaker) to mimic a humanoid behavior. The recorded microphones signals are enhanced by algorithms for dereverberation, ego-noise suppression, spatial filtering (beamforming or source separation) and post-filtering to improve the recognition rate of the subsequent automatic speech recognition (ASR) system. A system for acoustic echo control (AEC) allows the robot to listen to a person while speaking at the same time ("barge-in"). The recognized utterances of the ASR system are fed into a speech dialogue system which controls the robot's response to a speaker. A sound event detection system can help the dialogue system to react to acoustic events like a ringing bell.

The aim of this contribution is to provide an overview about some basic concepts for microphone array processing for robot audition and some recent advances. In Sec. 2, concepts for the placement of the robot microphones are presented. Algorithms for acoustic source localization and tracking are treated in Sec. 3. Approaches for ego-noise suppression and dereverberation are discussed in Sec. 4 whereas Sec. 5 treats spatial filtering and AEC for robot audition. The paper concludes with Sec. 6.

2. MICROPHONE ARRAY ARRANGEMENT

The design of a microphone array for robot audition can be based on two different paradigms. A first one is to consider a binaural system to mimic the auditory system of humans, e.g., [6, 7, 8]. A second one is to use as many microphones as technically possible and useful. For example, the commercially available robot NAO (version 5) of the manufacturer Softbank Robotics (formerly Aldebaran Robotics) contains 4 microphones in its head. A head array with 8 microphones is used for the humanoid robots SIG2, Robovie IIs and ASIMO [4]. A robot platform with 16 microphones has been considered in the

[1] `https://robot-ears.eu`

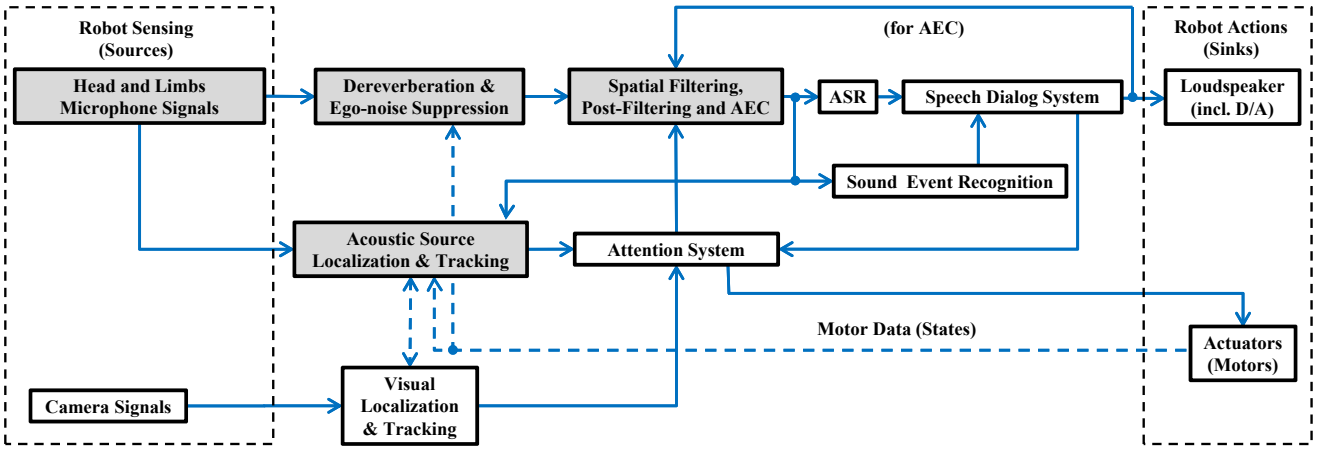

Fig. 1. *Block diagram of an overall system for robot audition.*

ROMEO project, cf., [9]. A circular array design with even 32 microphones for robot audition has been proposed in [10].

An important issue in the mechanical design of a robot head is to find the most suitable positions for the microphones. In [8], an approach to determining the optimal positions for a binaural microphone arrangement is proposed. The idea is to maximize the binaural cues, such as the interaural level difference (ILD) and interaural time difference (ITD), in dependence of the sensor positions to obtain the best possible localization performance. The ILD and ITD are expressed analytically by a spherical head model for the head-related transfer functions (HRTFs) of the robot head.

A framework to determine the most suitable positions for an arbitrary number of head microphones with respect to beamforming and DOA estimation is presented in [11]. It is based on the effective rank of a matrix composed of the generalized head-related transfer functions (GHRTFs). The optimal microphone positions can then be found by maximizing the effective rank of the GHRTF matrix for a given set of microphone and source positions. An extension of this concept has been developed to determine the optimal microphone positions of a spherical microphone array which maximizes the aliasing-free frequency range of the array [12, 13].

This new framework has also been used within the EARS project to construct a prototype head with 12 microphones for the robot NAO (shown in Fig. 2-a). The needed GHRTFs have been obtained by numerical simulation (cf., [11]) and areas, where a mounting of the microphones was not possible due to mechanical constraints, have been excluded for the numerical optimization.

The head microphone array might be extended by mounting microphones at the body and limbs of the robot termed as *robomorphic array* (Fig. 2-b). An additional benefit of this approach is that a higher array aperture can be realized than by using the head array microphones. If microphones are mounted at the robot limbs, the array aperture can be even varied by robot movements (provided that the robot still shows a natural behavior). This concept of the robomorphic array has been proposed for target signal extraction in [14]. The main idea is to run two "competing" blind signal extraction (BSE) algorithms [15] and to change the array aperture of the algorithm with the inferior signal extraction performance until its performance becomes superior and repeat this procedure continually.

A combination of head array and robomorphic array can also be exploited to improve the estimation of the DOA for a rotating head [16]. For the relatively small head of the NAO robot, the head array exhibits a relatively low estimation accuracy for frequencies below 1 kHz which can then be significantly improved by the use of a robomorphic array.

a) b)

Fig. 2. *a) Design drawing of the new 12-microphone prototype head, b) NAO robot with new prototype head and robomorphic array.*

3. ACOUSTIC SOURCE LOCALIZATION AND TRACKING

Effective robot audition requires awareness of the sound scene including the positions of sound sources and their trajectory over time. Estimates of source localization are needed, for example, to steer the beamformer and to track talkers. Time-varying acoustic maps can be used to capture this type of information. In acoustic localization, it is common that only bearing estimation can be obtained (DOA estimation), while range information is normally not available. A volumetric array comprising at least 4 microphones is required to identify a unique DOA in 2 dimensions (azimuth and inclination). A *spherical harmonics* (SH) representation of the sound field can be assumed for the almost spherical shape of a robot head, which suggests to perform the DOA estimation in the SH-domain. A method with low computational cost based on pseudointensity vectors (PIVs) [17, 18] is attractive given the limitations of robot embedded processing. This approach has been enhanced in [19], albeit with additional computational cost, to use subspace PIVs in order to obtain better performance and robustness.

Unlike many other applications of microphone arrays, robot-based arrays move. For DOA estimation, it is therefore necessary to account for the dynamics such as in the motion compensation approach of [20]. However, movement also enables to acquire additional spatial information and this can be exploited to enhance the DOA estimation performance in comparison to static sensing [21]. In addition to the movement of the robot, DOA estimation has to

account for the movement of the sound sources since talkers are rarely stationary. It is advantageous to employ tracking techniques that exploit models of source movement in order to improve on the raw output of a DOA estimator. This is challenging to achieve from acoustic sensing because of the lack of range information. Bearing-only source tracking has been developed, e.g., in [22] which exploits movement of the robot to estimate the location tracks as talkers move. Tracking is also advantageous in improving robustness to missing data and estimation variance. When the robot explores the acoustic environment, it needs to determine simultaneously its own location as well as a map of other sound sources in the vicinity. Techniques for *acoustic simultaneous localization and mapping* (A-SLAM) are proposed in [23], which localize the moving array and infer the missing source-sensor range from the estimated DOAs. DOA estimation accuracy is commonly degraded in reverberant environments due to acoustic reflections. The direct-path dominance test and the direct-path relative transfer function are exploited in the methods of [24, 25] that aim to base the DOA estimates mostly on direct path acoustic propagation rather than the acoustic reflections, thereby greatly improving robustness to the reverberation commonly encountered in use cases as for service robots.

If the target sources are in the field-of-view of the robot's cameras, audio-visual localization and tracking should be exploited, e.g., [26], which is beyond the scope of this paper.

4. EGO-NOISE REDUCTION AND DEREVERBERATION

The audio signals recorded by the microphones of the robot are usually not only distorted by external sources (room reverberation, background noise etc.), but also ego-noise caused by the actuators and CPU cooling fan of the robot, e.g., [2]. Thereby, the challenging task of suppressing the *non-stationary* actuator ego-noise is usually accomplished by exploiting information about the motor states and *a priori* knowledge about the specific structure of the noise sources using, e.g., a database with noise templates [27] or ego-noise prediction based on neural networks [28] where the actual enhancement is performed by spectral subtraction.

A multichannel approach, which considers also the phase information of the ego-noise, has been proposed in [29]. The actuator ego-noise is suppressed by a phase-optimized K-SVD algorithm where the needed dictionary is learned by sparse coding using multichannel ego-noise recordings. Ego-noise samples are modeled by a sparse combination of ego-noise prototype signals in the STFT-domain and capture the spectral as well as spatial characteristics of the current ego-noise sample. The evaluation of this approach for the NAO robot in [29] shows that a better speech quality and lower word error rate (WER) is achieved in comparison to related approaches based on non-negative matrix factorization (NMF) [30] or conventional K-SVD [31]. An extension of this approach in [32] uses nonlinear classifiers to associate the current motor state of the robot to relevant sets of entries in the learned dictionary. This approach achieves a significant reduction of the computational complexity in comparison to the original approach [29] while achieving at least the same noise reduction performance.

Besides ego-noise, room reverberation causes a significant degradation of the recorded audio signals and, hence, the ASR performance. In [33, 34], multichannel dereverberation is performed by MINT-filtering to enhance the performance of the subsequent signal separation by independent component analysis (ICA).

The almost spherical shape of a robot head suggests to perform the dereverberation in the SH-domain. In [35], the generalized weighted prediction error (GWPE) algorithm [36] for speech dereverberation is formulated in the SH-domain and offers computational savings over a conventional space-domain implementation when a high number of microphones is used.

5. SPATIAL FILTERING AND ACOUSTIC ECHO CANCELLATION

Spatial filtering for multichannel signal enhancement for robot audition can be realized by a data-dependent approach, e.g., [37, 38], a data-independent approach, e.g., [39] or a combination of both approaches, e.g., [9]. A data-dependent approach usually achieves a higher signal enhancement than a data-independent approach at the cost of a higher computational complexity. Moreover, the required statistics, e.g., covariance matrices, need to be estimated for highly nonstationary signals in the case of robot (head) movements.

A data-dependent approach for spatial sound source separation is given by the geometric source separation (GSS) [40]. Unlike the linearly constrained minimum variance (LCMV) beamformer, which minimizes the output power subject to a distortionless constraint for the target and additional constraints for interferers, GSS minimizes the cross-talk explicitly which leads to a faster adaptation [40]. An efficient realization of this approach is presented in [37] for robot audition (online GSS) as well as [38].

A recent, more general framework, which extends the LCMV concept to higher order statistics and uses ICA for a continuous estimation of noise and suppression of multiple interferers, has been proposed in [41, 42]. For the robot application, this approach can be implemented based on multiple two-channel BSS units [43] to allow for an extraction of multiple target sources [44].

A benefit of signal enhancement by data-independent fixed beamformers is their low computational complexity since the beamformer coefficients can be calculated in advance for different DOAs. However, the design of a beamformer is usually carried out by assuming a free-field propagation of sound waves, which is inappropriate for robot audition due to sound scattering effects at the robot head and torso. In [9], a minimum variance distortionless response (MVDR) beamformer design is proposed, where the HRTFs of a robot are used instead of a steering vector based on the free-field assumption. This HRTF-based beamformer is used as pre-processing for a subsequent blind source separation (BSS) system to reduce the reverberation and background noise. The evaluation of this approach reveals that this pre-processing step leads to a significant enhancement of the signal quality for the BSS [9].

In [39], the robust least-squares frequency-invariant (RLSFI) beamformer design of [45] has been extended by incorporating HRTFs of a robot head as steering vectors into the beamformer design to accounts for the sound scattering of a robot's head. An evaluation of this HRTF-based RLSFI beamformer design for the NAO robot head with five microphones has shown that a significantly better speech quality and lower WER can be achieved in comparison to the original free-field-based design as long as the HRTFs match the position of the target source [46]. An extension of the HRTF-based RLSFI beamformer design to the concept of polynomial beamforming is presented in [47], which allows for a flexible steering of the main beam without significant performance loss. In addition, the HRTF-based RLSFI beamformer design [39] has been extended such that the beamformer response can be controlled for all directions on a sphere surrounding the humanoid robot [48].

As suggested before, the almost spherical shape of a humanoid robot motivates to perform the beamforming in the SH-domain. The SH transformation of the sound pressure at the head microphone can be computed by using a boundary-element model for the robot head [49]. Based on this, well-known beamformers such as the maximum directivity beamformer or the delay-and-sum beamformer can

be implemented in the SH-domain [50]. To address the spatial aliasing problem for spherical arrays [12], a new general framework has been developed which can also be applied to robot heads [13].

The single-channel output signal of the spatial filtering system can be further enhanced by post-filtering. The needed noise power spectral density (PSD) is usually estimated by the input signals of the spatial filter. In [37] and [51], a post-filter is proposed whose filter weights are calculated by the MMSE amplitude estimator of [52]. The needed noise PSD is estimated by assuming that the transient components of the corrupting sources are caused by the leakage from other channels in the process of GSS. An evaluation on the SIG2 robot platform has revealed that this post-filtering approach achieves a significant reduction of the WER [51].

A humanoid robot needs a system for AEC such that it can listen to a person while speaking at the same time to allow for a so-called "barge-in". Most approaches for robot audition are based on a combination of spatial filtering and AEC; as already investigated in [53]. In [54], the AEC is performed on the input signals of a generalized sidelobe canceler (GSC) and the adaptation of the AEC filters is controlled by a double-talk detection, which considers the ratio of the PSDs of beamformer output and echo signal. In [33], the AEC is realized by means of ICA. Recently, it has been shown in [55] that a combination of GSC and AEC, where the AEC filter is operated in parallel to the interference canceler of the GSC according to [56], can also be successfully employed for robot audition.

6. CONCLUSIONS & OUTLOOK

The development of multichannel systems and algorithms for robot audition has received increased research interest in the last decades. The needed microphones are usually mounted to the robot head whose almost spherical shape motivates the use of SH-domain processing for spatial filtering and target source localization and tracking, if a large number of microphones is available. The optimal microphone positions can be found by numerical optimization (maximizing the effective rank of the GHRTF-matrix). The head microphone array might be extended by microphones integrated into the limbs and the body of the robot to increase the array aperture. A major challenge of this approach is to account for the varying sensor spacings due to robot movements.

Techniques for A-SLAM allow to localize the moving array and to infer the missing source-sensor range from the estimated DOAs. Such approaches have still a rather high computational complexity, but show promise for providing sophisticated acoustic awareness for robots in the future. The ego-noise reduction is usually performed by exploiting a priori knowledge about the specific structure of the noise sources and by incorporating information about the motor states. Recent works suggest that it is beneficial to consider also the relative phase of the ego-noise components in the multichannel recordings. Advanced techniques for dereverberation and spatial filtering can also be employed for robot audition, where the HRTFs of the robot head should be considered in the design of such systems. The adaptive AEC, which is needed to allow for a "barge-in", is usually designed jointly with the spatial filtering to ensure a fast convergence.

A promising direction for future robot audition systems is to benefit from external sensors, which may be provided, e.g., by all kinds of voice communication devices, smart home environments or other robots.

Acknowledgment

The authors would like to thank Hendrik Barfuss, Alexander Schmidt, Christine Evers and all other coworkers in the EARS project for their contributions which form the background of this paper.

7. REFERENCES

[1] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 5610–5614.

[2] H. W. Löllmann, H. Barfuss, A. Deleforge, and W. Kellermann, "Challenges in acoustic signal enhancement for human-robot communication," in *ITG Conf. on Speech Communication*, Erlangen, Germany, Sept. 2014, pp. 1–4.

[3] K. Nakadai, G. Ince, K. Nakamura, and H. Nakajima, "Robot audition for dynamic environments," in *Intl. Conf on Signal Processing, Communications and Computing (ICSPCC)*, Hong Kong, China, Aug. 2012, pp. 125–130.

[4] H. G. Okuno, T. Ogata, and K. Komatani, "Robot audition from the viewpoint of computational auditory scene analysis," in *Intl. Conf. on Informatics Education and Research for Knowledge-Circulating Society (icks)*, Kyoto, Japan, Jan. 2008, pp. 35–40.

[5] G. Schillaci, S. Bodiroža, and V. V. Hafner, "Evaluating the effect of saliency detection and attention manipulation in human-robot," *Intl. Journal of Social Robotics, Springer*, vol. 5, no. 1, pp. 139–152, 2013.

[6] R. Liu and Y. Wang, "Azimuthal source localization using interaural coherence in a robotic dog: Modeling and application," *Robotica*, vol. 28, pp. 1013–1020, 2010, Cambridge University Press.

[7] S. Argentieri, A. Portello, M. Bernard, P. Danés, and B. Gas, "Binaural systems in robotics," in *The Technology of Binaural Listening*, J. Blauert, Ed., Modern Acoustics and Signal Processing, pp. 225–253. Springer, 2013.

[8] A. Skaf and P. Danés, "Optimal positioning of a binaural sensor on a humanoid head for sound source localization," in *IEEE-RAS Intl. Conf. on Humanoid Robots*, Bled, Slovenia, Oct. 2011, pp. 165–170.

[9] M. Maazaoui, K. Abed-Meraim, and Y. Grenier, "Adaptive blind source separation with HRTFs beamforming preprocessing," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Hoboken, NJ, USA, June 2012, pp. 269–272.

[10] Y. Tamai, S. Kagami, Y. Amemiya, Y. Sasaki, H. Mizoguchi, and T. Takano, "Circular microphone array for robot's audition," in *IEEE Sensors 2004*, Valencia, Spain, Oct. 2004, vol. 2, pp. 565–570.

[11] V. Tourbabin and B. Rafaely, "Theoretical framework for the optimization of microphone array configuration for humanoid robot audition," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1803–1814, Dec. 2014.

[12] D. L. Alon and B. Rafaely., "Beamforming with optimal aliasing cancellation in spherical microphone arrays," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 196–210, Jan. 2016.

[13] V. Tourbabin and B. Rafaely, "Optimal design of microphone array for humanoid-robot audition," in *Israeli Conf. on Robotics (ICR)*, Herzliya, Israel, Mar. 2016, (abstract).

[14] H. Barfuss and W. Kellermann, "An adaptive microphone array topology for target signal extraction with humanoid robots," in *Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Antibes, France, Sept. 2014, pp. 16–20.

[15] Y. Zheng, K. Reindl, and W. Kellermann, "BSS for improved interference estimation for blind speech signal extraction with two microphones," in *Intl. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Aruba, Dutch Antilles, Dec. 2009, pp. 253–256.

[16] V. Tourbabin, H. Barfuss, B. Rafaely, and W. Kellermann, "Enhanced robot audition by dynamic acoustic sensing in moving humanoids," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 5625–5629.

[17] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 442–446.

[18] A. H. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test," in *European Signal Processing Conf. (EUSIPCO)*, Nice, France, Aug. 2015, pp. 2296–2300.

[19] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudo-intensity vectors," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 178 –192, Jan. 2017.

[20] V. Tourbabin and B. Rafaely, "Speaker localization by humanoid robots in reverberant environments," in *IEEE Conv. of Electrical and Electronics Engineers in Israel (IEEEI)*, Eilat, Dec. 2014, pp. 1–5.

[21] V. Tourbabin and B. Rafaely, "Utilizing motion in humanoid robots to enhance spatial information recorded by microphone arrays," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Nancy, France, May 2014, pp. 147–151.

[22] C. Evers, A. H. Moore, P. A. Naylor, J. Sheaffer, and B. Rafaely, "Bearing-only acoustic tracking of moving speakers for robot audition," in *IEEE Intl. Conf. Digital Signal Processing (DSP)*, Singapore, July 2015, pp. 1206–1210.

[23] C. Evers, Alastair, and P. A. Naylor, "Acoustic simultaneous localisation and mapping (a-SLAM) of a moving microphone array and its surrounding speakers," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 6–10.

[24] X. Li, L. Girin, R. Horaud, and S. Gannot, "Local relative transfer function for sound source localization," in *European Signal Processing Conf. (EUSIPCO)*, Nice, France, Aug. 2015, pp. 399–403.

[25] B. Rafaely and D. Kolossa, "Speaker localization in reverberant rooms based on direct path dominance test statistics," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, Mar. 2017.

[26] B. Bayram and G. Ince, "Audio-visual human tracking for active robot perception," in *Signal Processing and Communications Applications Conf. (SIU)*, Malatya, Turkey, May 2015, pp. 1264–1267.

[27] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura, "Ego noise suppression of a robot using template subtraction," in *IEEE/RAS Intl. Conf. on Intelligent Robots and Systems (IROS)*, St. Louis, MO, USA, Oct. 2009, pp. 199–204.

[28] A. Ito, T. Kanayama, M. Suzuki, and S. Makino, "Internal noise suppression for speech recognition by small robots," in *European Conf. on Speech Communication and Technology (Interspeech)*, Lisbon, Portugal, Sept. 2005, pp. 2685–2688.

[29] A. Deleforge and W. Kellerman, "Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 355–359.

[30] Y. Li and A. Ngom, "Versatile sparse matrix factorization and its applications in high-dimensional biological data analysis," *Pattern Recognition in Bioinformatics*, pp. 91–101, 2013, Springer.

[31] M. Aharon, M. Elad, and A. Bruckstein, "K -SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[32] A. Schmidt, A. Deleforge, and W. Kellermann, "Ego-noise reduction using a motor data-guided multichannel dictionary," in *IEEE/RAS Intl. Conf. on Intelligent Robots and Systems (IROS)*, Daejon, Korea, Oct. 2016, pp. 1281–1286.

[33] R. Takeda, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno, "ICA-based efficient blind dereverberation and echo cancellation method for barge-in-able robot audition," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3677–3680.

[34] R. Takeda, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno, "Speedup and performance improvement of ICA-based robot audition by parallel and resampling-based block-wise processing," in *IEEE/RAS Intl. Conf. on Intelligent Robots and Systems (IROS)*, Oct. 2010, pp. 1949–1956.

[35] A. H. Moore and P. A. Naylor, "Linear prediction based dereverberation for spherical microphone arrays," in *Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, Sept. 2016, pp. 1–5.

[36] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[37] J. M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *IEEE/RAS Intl. Conf. on Intelligent Robots and Systems (IROS)*, Sendai, Japan, Sept. 2004, vol. 3, pp. 2123–2128.

[38] K. Nakadai, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Sound source separation of moving speakers for robot audition," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3685–3688.

[39] H. Barfuss, C. Hümmer, G. Lamani, A. Schwarz, and W. Kellermann, "HRTF-based robust least-squares frequency-invariant beamforming," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2015, pp. 1–5.

[40] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, Sept. 2002.

[41] K. Reindl, S. Markovich-Golan, H. Barfuss, S. Gannot, and W. Kellermann, "Geometrically constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.

[42] K. Reindl, S. Meier, H. Barfuss, and W. Kellermann, "Minimum mutual information-based linearly constrained broadband signal extraction," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1096–1108, June 2014.

[43] C. Anderson, S. Meier, W. Kellermann, P. Teal, and M. Poletti, "A GPU-accelerated real-time implementation of TRINICON-BSS for multiple separation units," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Nancy, France, May 2014, pp. 102–106.

[44] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Combined LCMV-TRINICON beamforming for separating multiple speech sources in noisy and reverberant environments," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 320 – 332, Feb. 2016.

[45] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 77–80.

[46] H. Barfuss and W. Kellermann, "On the impact of localization errors on HRTF-based robust least-squares beamforming," in *DAGA 2016*, Aachen, Germany, Mar. 2016, pp. 1072–1075.

[47] H. Barfuss, M. Müglich, and W. Kellermann, "HRTF-based robust least-squares frequency-invariant polynomial beamforming," in *Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, Sept. 2016, pp. 1–5.

[48] H. Barfuss, M. Bürger, J. Podschus, and W. Kellermann, "HRTF-based two-dimensional robust least-squares frequency-invariant beamformer design for robot audition," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, CA, USA, Oct. 2017.

[49] V. Tourbabin and B. Rafaely, "Direction of arrival estimation using microphone array processing for moving humanoid robots," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2046–2058, Nov. 2015.

[50] B. Rafaely, "Phase-mode versus delay-and-sum spherical microphone array processing," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 713–716, Oct. 2005.

[51] S. Yamamoto, J.-M. Valin, K. Nakadai, J. Rouat, F. Michaud, T. Ogata, and H. G. Okuno, "Enhanced robot speech recognition based on microphone array source separation and missing feature theory," in *Intl. Conf. on Robotics and Automation (ICRA)*, Barcelona, Spain, Apr. 2005, pp. 1477–1482.

[52] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processsing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[53] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Munich, Germany, Apr. 1997, vol. 1, pp. 219–222.

[54] J. Beh, T. Lee, I. Lee, H. Kim, S. Ahn, and H. Ko, "Combining acoustic echo cancellation and adaptive beamforming for achieving robust speech interface in mobile robot," in *IEEE/RAS Intl. Conf. on Intelligent Robots and Systems (IROS)*, Nice, France, Sept. 2008.

[55] A. El-Rayyes, H. W. Löllmann, C. Hofmann, and W. Kellermann, "Acoustic echo control for humanoid robots," in *DAGA 2016*, Aachen, Germany, Mar. 2016, pp. 1–4.

[56] W. Herbordt, W. Kellerman, and S. Nakamura, "Joint optimization of LCMV beamforming and acoustic echo cancellation," in *European Signal Processing Conf. (EUSIPCO)*, Lisbon, Portugal, Sept. 2004, pp. 2003–2006.